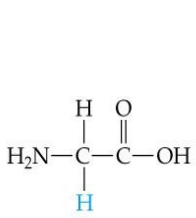
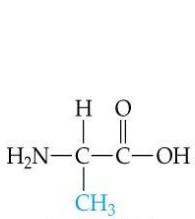
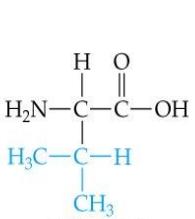
Bioinformatics
CS300
Substitution Matrices and Protein Alignments
(Chap 4 and 5 in textbook)

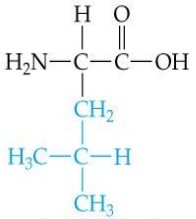
Week9, Deck 1
Fall 2022
Oliver BONHAM-CARTER

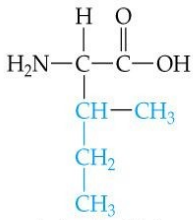
Amino Acids

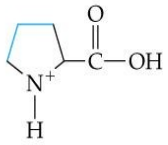

Glycine (Gly)

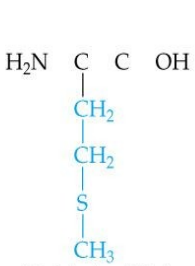

Alanine (Ala)

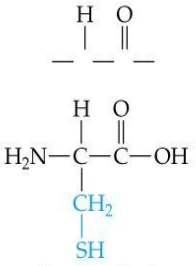

Valine (Val)

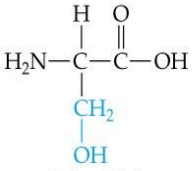

Leucine (Leu)

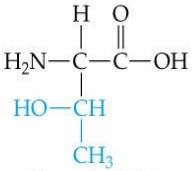

Isoleucine (Ile)

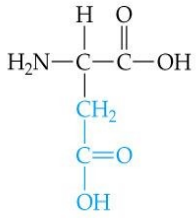

Proline (Pro)

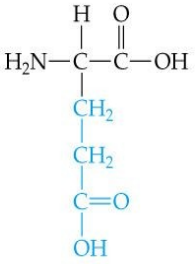

Methionine (Met)

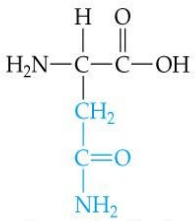

Cysteine (Cys)

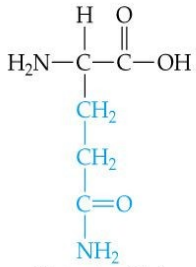

Serine (Ser)

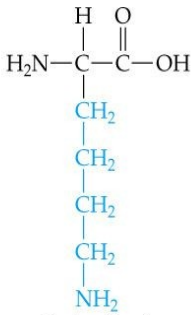

Threonine (Thr)

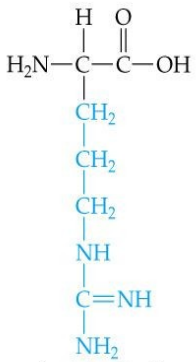

Aspartic acid (Asp)

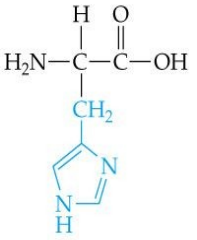

Glutamic acid (Glu)

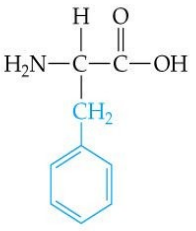

Asparagine (Asn)

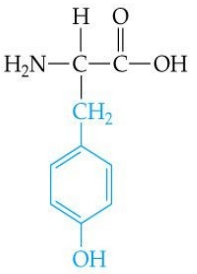

Glutamine (Glu)

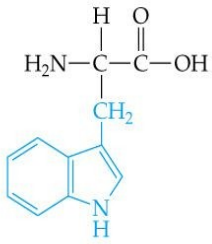

Lysine (Lys)


Arginine (Arg)


Histidine (His)


Phenylalanine (Phe)

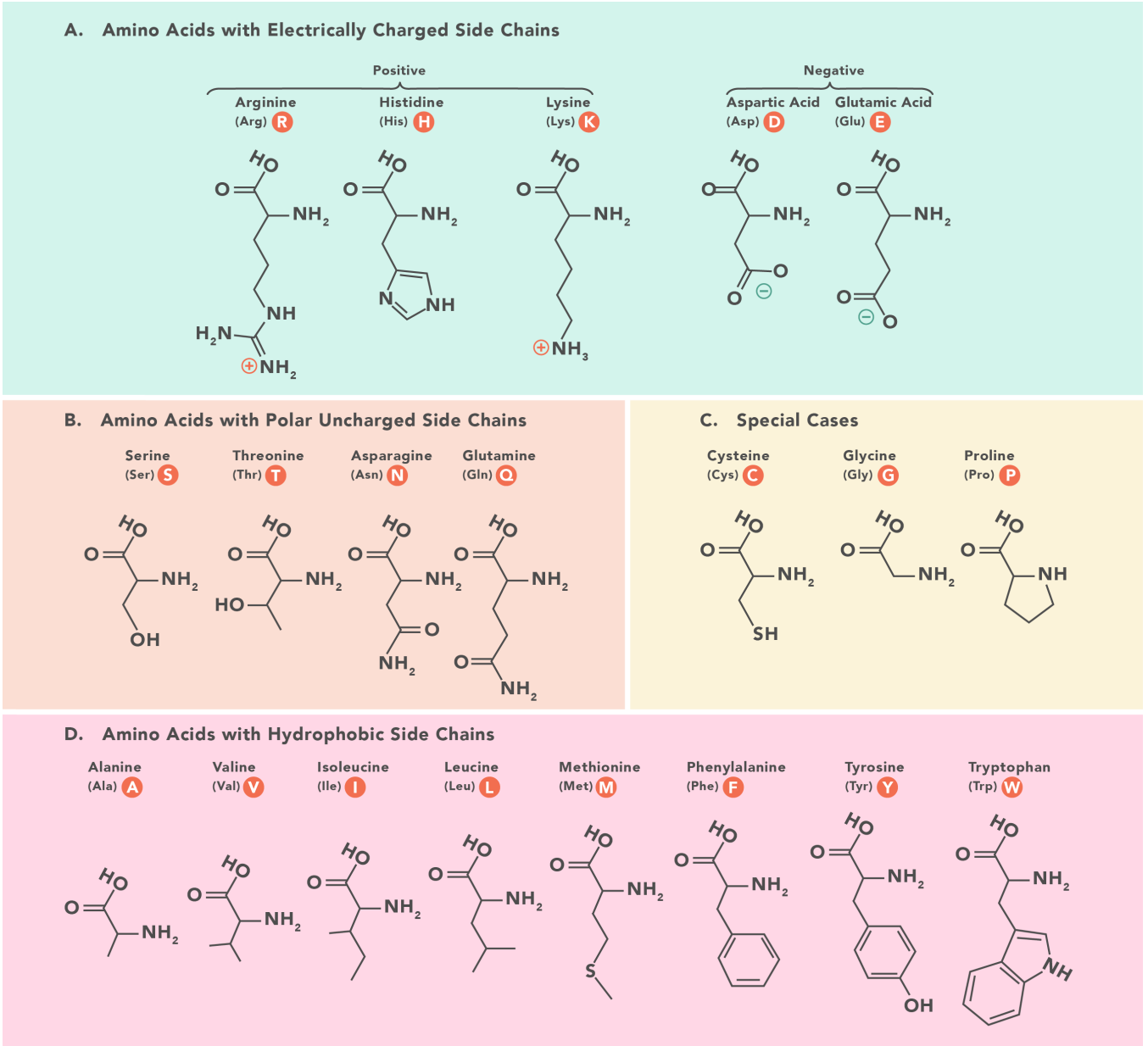

Tyrosine (Tyr)


Tryptophan (Trp)

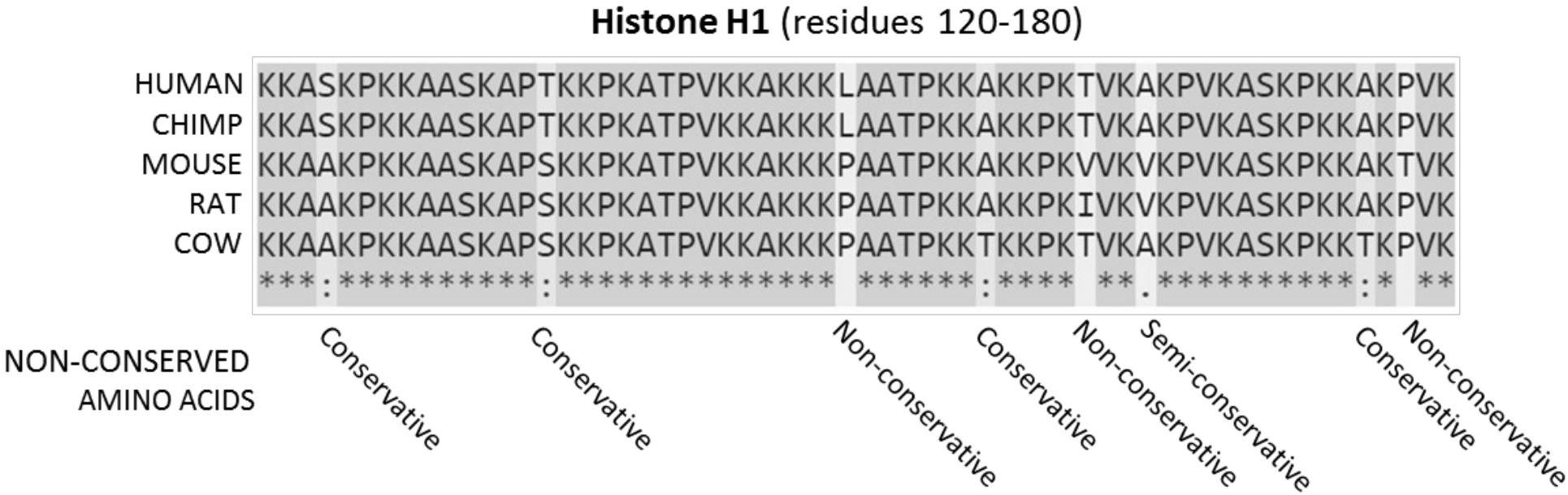
Physicochemical properties

(Physics + chemistry)

- Polar vs nonpolar
- Hydrophobic vs hydrophilic
- Positive electric charge vs negative electric charge
- Basic vs Acidic

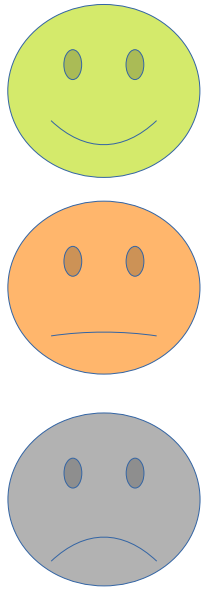


Protein Amino Acid Replacements



Generally, replacements are ...

- **Conservative**: a change to an amino acid with similar physio-chemical properties; a smaller effect on function than non-conservative replacements.
- **Semi-conservative**: Minor changes that persist, depending on evolutionary conditions
- **Non-conservative:** Changes that are likely to be edited out by evolutionary pressures due to their deleterious effects



Quantification of Traits

- Could we quantify sequence by physicochemical properties? (yes!)

Table 5.1 Hydrophobicity values for the 20 amino acids. A more positive value represents a more hydrophobic amino acid.

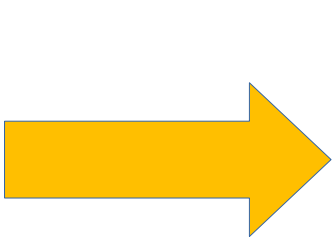
Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity
D	−3.5	Y	−1.3	I	4.5
K	−3.9	N	−3.5	C	2.5
H	−3.2	L	3.8	A	1.8
T	−0.7	E	−3.5	S	−0.8
V	4.2	R	−4.5	G	−0.4
F	2.8	W	−0.9	P	−1.6
M	1.9	Q	−3.5		

“Randomness” in Protein?

- **Thought experiment**: Can protein really be all that *random* in nature?
- We can generate random protein sequences, but are they found in nature?
- Make your own random protein sequence!
  https://www.bioinformatics.org/sms2/random_protein.html

Random Protein Sequence results

```
>random sequence 1 consisting of 1000 residues.
ISRYVYIYVQQNMGWFTHHPCQHCITFAKMCFRWNWAIGPEWCQLRWPTYMGVFWKWAIF
PSHRHMTLKTFDYKPLQFIIAQEEYPLLEFSMGQGKLSDKEFYIHCHFPICWDIWTSFED
VKKWQTQYEKIAYRNVYQQTMDDDWLLKNNWDFTYIMLNCVHILGQHGAGHDCMATYQAH
CDSKTGNTLHYFDRMCQDKMPKANHCPWEHEYMGPVAGLSDEMKIQKHNHSFRGTMSEHG
THMHRCMANLLDPYVMQECLDAIYFDKPGTRFPRYKLVCNIYGHYWHAGHFVPWDGPARE
KAEQVLNNYAVFGSKDSACQGQKSHFDPNTCCEVQNIPPQMDLYYCNRGFRQLMQTCDMK
NMNDSWMQAHFMWQYPCLKSTRVLQNNALSLWTTIMDVQYVMAPRPAPYPMWIGCILKLI
HMMELWFQEPCAVQWCYMIEGLRVHGSQAHHKFEYQTYAIAGQHGWYWPPTMQSIESGAS
DPNYDHLDSEHLNEEMQGFVCYFLQYHAGKFNTSGTLMFRAQREKLMKAKIWHIATLRKL
EKCQAKGYLMMKLGWGAMVSGHRINNHKKYVDKAVNDPPAPLHITHTPTCHYYAHNFRKH
AQIPGELMILEVGLAQEAYACYRRMYCDIIGTWYRFECFHNLMKSDMPDVSLFEKWHYEE
CDEPLQTFMPPFSCYQWDHIWHKMDSEQMCDRLRNCAIDLFFMDFWCQYTPCNAYMPQRW
SRFVEAERQDYAREGVHEPTSYWTVQHFSNFTLLRHQHDVSWPKWMMFKHWYVCSGFDFK
ALITNTAVTWNVKIFCFKWCIHKSDFAQLARFFPWGFNRMTSPRQNQMCVVYCQRAFREI
MRTFQPCSKHVWDNSVLAEGRAKDWHGMYLTARTFYEYTHRGSFWFKCAIHIESWDLRDL
QDCIMRVLDTRRDDASSYFLIFLEFFAHPEVSCFDVFKHFIILTVFMHGQCAVPDVHDEA
WMWPIHIEYQFPNSAQWAIIFVANCVNTPTKWALEVQFKP
```



Random Protein Sequence

Protein Blast: where is this sequence found in nature?

Protein: Statistical Interests

- Statistically Significant: With a larger protein "alphabet" (20 amino acids), it is much less likely to get matches by chance.
- Amino acid changes are not equally harmful to protein structure

Chance of "**M**ethioine-**L**eucine-**S**erine" occurring "randomly"

P(*M*) * P(*L*) * P(*S*)
= (1/20) * (1/20) * (1/20)
= (1/20)^3
= 0.000125, or 0.0125 percent

Longer words are even less likely!

Origin: Probability of a Single Protein Forming by Chance:
https://www.youtube.com/watch?v=W1_KEVaCyaA

Scoring Amino Acid Substitutions

Better to study evolution of real proteins from closely related organisms

Minimizes likelihood that an observed difference represents a series of more than one individual mutation

Species A – Ala Species B – Ile

No intermediate mutations?

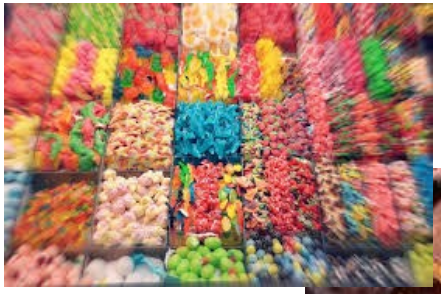
Ala --> Ile : 1 mutation

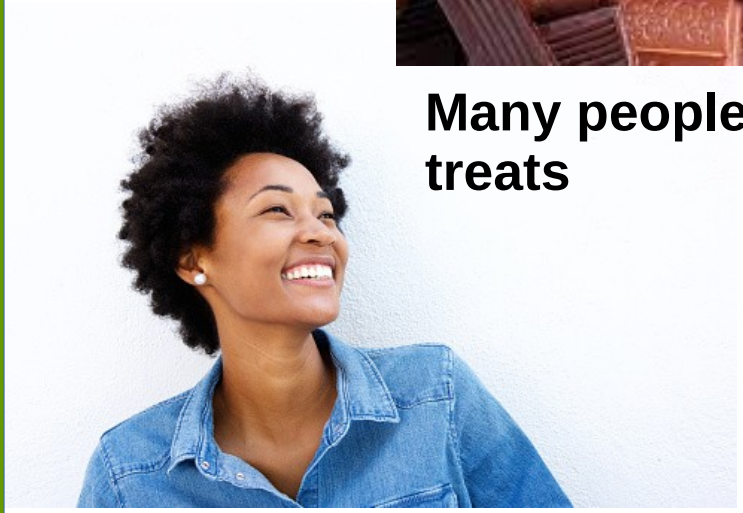
Ala --> Pro --> Ser --> Ile : 3 mutations

A few intermediate mutations?

Some Events Are More *Likely*

Watching sports together rather than separately





Many people enjoy sweet treats

Dogs chase cats (mostly)

A Model of Evolutionary Change in Proteins, Dayhoff *et al.*, 1978

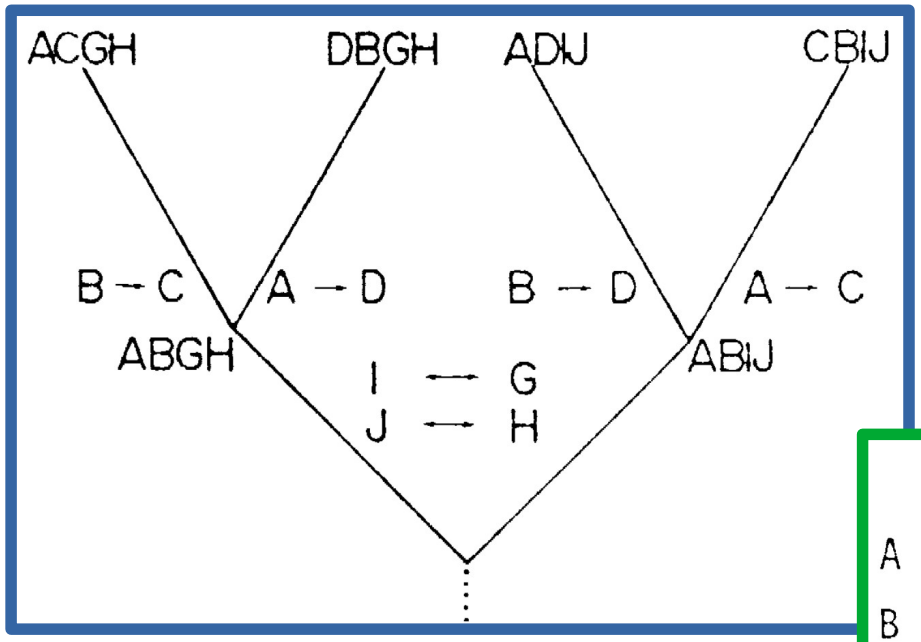
Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

- M_{ij} - the probability of a mutation replacing amino *i* with animo *j*
- f_j - the frequency of amino acid *j* in a large set of sequences

A Model of Evolutionary Change in Proteins, Dayhoff *et al.*, 1978



Tree of observed amino acid changes

Frequencies of amino acid changes added to matrix

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

A Model of Evolutionary Change in Proteins, Dayhoff *et al.*, 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

- Greater positive for likely (**conservative**) substitutions
- Greater negative for unlikely (**non-conservative**) substitutions
- Multiplied by 10 and rounded to nearest integer

A Model of Evolutionary Change in Proteins, Dayhoff *et al.*, 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

What's the observed frequency of one amino acid being replaced for another (with the likelihood of finding the amino acid *j* by chance)?

Logs-Odds Ratio

- **Log-Odds(X,Y) >0: x is likely to be replaced by Y in nature**
- **Log-Odds(X,Y) <0: x not likely to be replaced with Y in nature**
- **Log-Odds(X,Y) = 0: replacement more likely to occur by chance.**

The PAM Matrix

- PAM matrices are used as substitution matrices to score sequence alignments for proteins.
- Each entry in a PAM matrix indicates the likelihood of the amino acid of that row being replaced with the amino acid of that column through a series of one or more point accepted mutations during a specified evolutionary interval, rather than these two amino acids being aligned due to chance.
- Different PAM matrices correspond to different lengths of time in the evolution of the protein sequence.

Ref:
https://en.wikipedia.org/wiki/Point_accepted_mutation

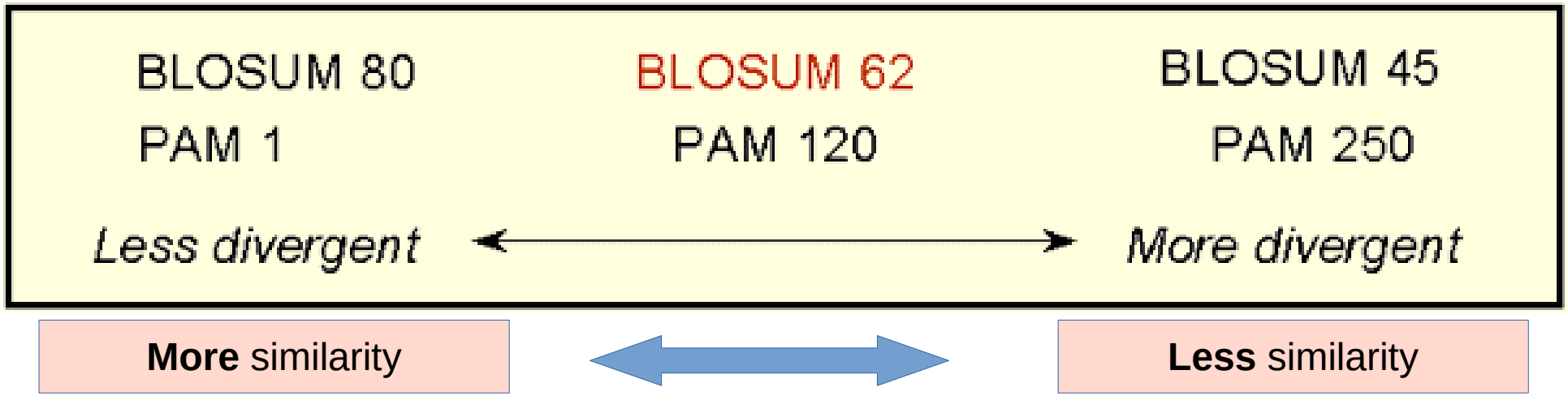
The PAM Matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	2																			
Arg	R	–1	5																		
Asn	N	0	0	3																	
Asp	D	0	–1	2	5																
Cys	C	–1	–1	–1	–3	11															
Gln	Q	–1	2	0	1	–3	5														
Glu	E	–1	0	1	4	–4	2	5													
Gly	G	1	0	0	1	–1	–1	0	5												
His	H	–2	2	1	0	0	2	0	–2	6											
Ile	I	0	–3	–2	–3	–2	–3	–3	–3	–3	4										
Leu	L	–1	–3	–3	–4	–3	–2	–4	–4	–2	2	5									
Lys	K	–1	4	1	0	–3	2	1	–1	1	–3	–3	5								
Met	M	–1	–2	–2	–3	–2	–2	3	3	–2	3	3	–2	6							
Phe	F	–3	–4	–3	–5	0	–4	–5	–5	0	0	2	–5	0	8						
Pro	P	1	–1	–1	–2	–2	0	–2	–1	0	–2	0	–2	–2	–3	6					
Ser	S	1	–1	1	0	1	–1	–1	1	–1	–1	–2	–1	–1	–2	1	2				
Thr	T	2	–1	1	–1	–1	–1	–1	–1	–1	1	–1	–1	0	–2	1	1	2			
Trp	W	–4	0	–5	–5	1	–3	–5	–2	–3	–4	–2	–3	–3	–1	–4	–3	–4	15		
Tyr	Y	–3	–2	–1	–2	2	–2	–4	–4	4	–2	–1	–3	–2	5	–3	–1	–3	0	9	
Val	V	1	–3	–2	–2	–2	–3	–2	–2	–3	4	2	–3	2	0	–1	–1	0	–3	–3	4

The probability calculations for Substitutions have been done for you!

PAM Matrices

- **P**oint **A**ccepted **M**utation
- Family of matrices PAM 1, PAM 80, PAM 120, PAM 250
- The number of PAM matrix (i.e., the '***n***' in PAM **# *n***) represents the evolutionary distance between the sequences on which the matrix is based



BLOSUM matrix
Heinkoff and Heinkoff, 1992

- **BLOcks SUbstition Matrix** - Blocks of local alignments

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

- p_{ij} - probability j replacing i
- q_i and q_j - probabilities of finding the amino acids i and j in any protein sequence
- λ - scaling factor, set such that the matrix contains easily computable integer values.

- BLOSUM # - # = minimum % similarity of sequences compared

BLOSUM matrix
Heinkoff and Heinkoff, 1992

- Sequence alignment of proteins
- matrices are used to score alignments between evolutionarily divergent protein sequences
- They are based on local alignments of very conserved regions of protein families (that do not have gaps in the sequence alignment)
- Relative frequencies of amino acids and their substitution probabilities calculated
- Log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids
- Note: BLOSUM matrices based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.

The BLOSUM Matrix

	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
A	0	1	0	4																	A
G	-3	0	-2	0	6																G
P	-3	-1	-1	-1	-2	7															P
D	-3	0	-1	-2	-1	-1	6														D
E	-4	0	-1	-1	-2	-1	2	5													E
Q	-3	0	-1	-1	-2	-1	0	2	5												Q
N	-3	1	0	-2	0	-2	1	0	0	6											N
H	-3	-1	-2	-2	-2	-2	-1	0	0	1	8										H
R	-3	-1	-1	-1	-2	-2	-2	0	1	0	0	5									R
K	-3	0	-1	-1	-2	-1	-1	1	1	0	-1	2	5								K
M	-1	-1	-1	-1	-3	-2	-3	-2	0	-2	-2	-1	-1	5							M
I	-1	-2	-1	-1	-4	-3	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-1	-4	-3	-4	-3	-2	-3	-3	-2	-2	2	2	4					L
V	-1	-2	0	0	-3	-2	-3	-2	-2	-3	-3	-3	-2	1	3	1	4				V
W	-2	-3	-2	-3	-2	-4	-4	-3	-2	-4	-2	-3	-3	-1	-3	-2	-3	11			W
Y	-2	-2	-2	-2	-3	-3	-3	-2	-1	-2	2	-2	-2	-1	-1	-1	-1	2	7		Y
F	-2	-2	-2	-2	-3	-4	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	1	3	6	F
	C	S	T	A	G	P	D	E	Q	N	H	R	K	M	I	L	V	W	Y	F	

PAM vs BLOSUM

- **General Use**
  - PAM 120
  - BLOSUM 62*
- **Closely Related Species**
  - PAM 60
  - BLOSUM 80
- **Distantly Related Species**
  - PAM 250
  - BLOSUM 45

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

*BLOSUM 62 – used by BLAST – computed by choosing blocks of local alignments more than 62% identical

** BLOSUM matrices are gradually replacing PAM matrices thanks to advanced data analysis for calculating probabilities of substitutions

In common: PAM and BLOSUM

PAM	BLOSUM
To compare closely related sequences, PAM matrices with lower numbers are created.	To compare closely related sequences, BLOSUM matrices with higher numbers are created.
To compare distantly related proteins, PAM matrices with high numbers are created.	To compare distantly related proteins, BLOSUM matrices with low numbers are created.

- **The two result in the same scoring outcome, but use differing methodologies.**

https://en.wikipedia.org/wiki/BLOSUM#The_relationship_between_PAM_and_BLOSUM

Differences: PAM and BLOSUM

PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.[19]

Related sequences or closely related sequences

- BLOSUM looks directly at mutations in motifs of related sequences
- PAM's extrapolate evolutionary information is based on closely related sequences

https://en.wikipedia.org/wiki/BLOSUM#The_relationship_between_PAM_and_BLOSUM

Differences: PAM and BLOSUM

PAM	BLOSUM
PAM matrices are used to score alignments between closely related protein sequences.	BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences.
Based on global alignments	Based on local alignments
Alignments have high similarity than BLOSUM alignments	Alignments have low similarity than PAM alignments
Mutations in global alignments are vey significant	based on highly conserved stretches of alignments
Higher numbers in the PAM matrix naming denotes greater evolutionary distance	Higher numbers in the BLOSUM matrix naming denotes higher sequence similarity and smaller evolutionary distance
Example: PAM 250 is used for more distant sequences than PAM 120	Example: BLOSUM 80is used for closely related sequences than BLOSUM 62

https://www.majordifferences.com/2014/02/difference-between-pam-and-blosum-matrix_1.html

Blast Subst Matrices

Substitution Matrices: BLOSUM
https://www.youtube.com/watch?v=0_66UK-439M

BLAST 5 BLOSUM62
https://www.youtube.com/watch?v=njva17LwhsE

BLAST substitution matrices
https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html